A NOVEL APPROACH OF CLASSIFICATION TECHNIQUES FOR CLIR

Archana.M and Dr. Sumithra Devi K.A[2]

[1]Asst.Professor, Dept of MCA, RV College of Engineering, Bangalore-59
E-mail: archanams_m@yahoo.com

[2]Director, Dept of MCA, RV College of Engineering, Bangalore-59
E-mail: sumithraka@gmail.com

ABSTRACT

Recent and continuing advances in online information systems are creating many opportunities and also new problems in information retrieval. Gathering the information in different natural language is the most difficult task, which often requires huge resources. Cross-language information retrieval (CLIR) is the retrieval of information for a query written in the native language. This paper deals with various classification techniques that can be used for solving the problems encountered in CLIR.

KEYWORDS

Information Retrieval, Cross Language Information Retrieval (CLIR), Data Mining

1. INTRODUCTION

In conventional world, information retrieval was mainly concerned with indexing the terms and search for the useful documents. Nowadays, research in IR includes modeling, document classification, categorization, search engines, user interfaces, data visualization, information filtering, natural language processing or query language and systems architecture. Also from the digital resource perspective, IR research includes text mining, multimedia retrieval, and digital libraries. The representation of IR system in different principles is given below:

	Finding answers and information that already exist in a system		Creating answers and new information by analysis and inference – based on query
	Search by navigation (following links, as in a subject directory and the Web generally)	Search by query (as in Google)	
Unstructured information (text, images, sound)	**Hypermedia systems** (Many small units, such as paragraphs and single images, tied together by links)	**IR systems** (Often dealing with whole documents, such as books and journal articles)	
Structured information		**Database management systems (DBMS)**	**Data analysis systems Expert systems**

Figure1: Information Retrieval System [14]

D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 222–230, 2011.

DOI: 10.5121/csit.2011.1221

The rapid growth of communication technologies has immense impact on information retrieval (IR) technique, which has allowed people worldwide to access previously unavailable information. With these advances, however, it has become clear that there is a growing need for retrieval of information in many languages. IR helps the user get useful information from digital resources including digital libraries, WWW and documents.

Some of the problems encountered while retrieving the document that gave rise to the cross language information retrieval are [1]:

i. The collection of document in different languages, where query formulation for each language would be extremely inefficient.
ii. Documents that contains text in different language (more than one language).
iii. User is not able to write the query apart from the native language, but able to make use of documents retrieved in different language that contains images or names not requiring much of the efforts.

Cross-Language information retrieval would be very useful in the fields of research, where lot of information can be accessed that are present in different language related to the required query. The main goal of cross language information retrieval (CLIR) is the growing need for exploring the document in the foreign languages. It tries to identify relevant documents in different language from that of the query. Simple knowledge structures such as bilingual term lists have proven to be a useful basis for bridging the language gap. Since the queries and documents are in different language, they have to be translated before they are matched.

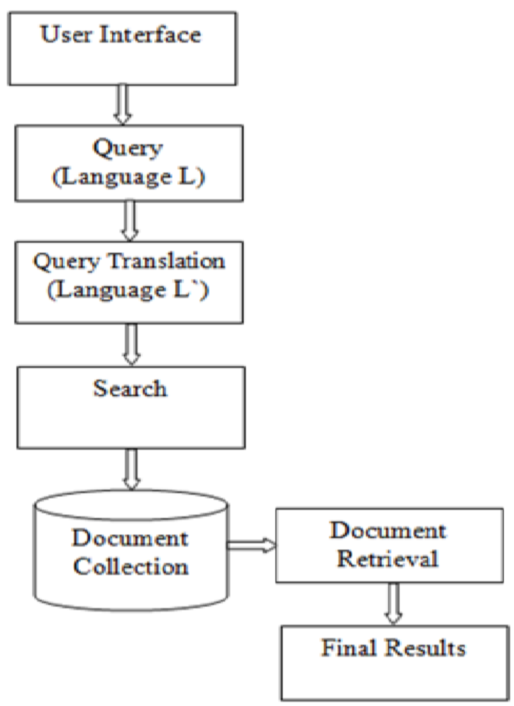

Figure2: Cross-Language information retrieval System

Knowledge Creation (Data mining) is a process of automatically discovering useful information in the large data set. Data mining techniques are deployed to search large databases in order to find useful patterns. It is classified into two types:

i. Direct – target field in explained in this method. The best examples of this method are classification, estimation and predication.
ii. Undirected – without the particular target field, this method try to fine the pattern or similarities among the groups and best examples are affinity

2. BASIC APPROACH FOR CLIR

Some of the basic approaches to CLIR are Machine translation, controlled vocabulary and dictionary-based approach

2.1 Machine Translation

One of the common methodologies to CLIR is the use of machine translation (MT), in which document or query can be translated automatically. This is one of the simplest and useful because the query can be translated from the language of the source to another language for search and translate the result back into the source language for viewing. The only snag with MT system is that it often makes translation errors because of missing information in the term index or ambiguous definitions. It can only produces high quality translations for specific domains, such as those containing specific technical terminology, possibly because semantic accuracy suffers when insufficient domain knowledge is incorporated into a translation system.

2.2 Controlled Vocabulary

The controlled vocabulary approach has been most assertive and efficient in the long run. A controlled vocabulary is a way to insert an interpretive layer of semantics between the term entered by the user and the underlying database to better represent the original intention of the terms of the user. CLIR a multilingual thesaurus of some sort is created to hold a list of descriptors for each document in a collection and the semantic relations between them, and each term in the thesaurus must be translated for each language involved. The descriptors can be added to the thesaurus manually or automatically if the system can learn from previous indexing which terms are likely to be important. The hitch in this approach is that a query must be generated using only vocabulary from the thesaurus, in which case it may be difficult to search for specific terms that are not included. The larger the size of the vocabulary in the thesaurus, the less effective it becomes.

2.3 Dictionaries-Based

In dictionary based CLIR each term (semantic unit, single word or a phrase) in the user query is looked up in the machine-readable bilingual dictionary. Some form of ambiguity resolution or equivalent selection is applied to pick the best translation of that term from the list given by the dictionary. This translation is then added to the document language semantic mapping of the bag of words query. This document language semantic mapping is then matching against the document collection as if it had been directly derived from the initial user request. DB for CLIR can be divided into four logical steps [12]:

i. Pre translation query modification
ii. Dictionary lookup

iii. Equivalent selection and weighting
iv. Post-translation query translation modification

The advantage of using a simple bilingual dictionary to translate query terms is that, dictionaries and wordlists covering a wide range of subject areas and language pairs are readily available. In addition, the time needed to implement and set up a dictionary-based system from a printed or electronic source is considerably less comparative to an MT engine for a new language pair. A machine-readable bilingual dictionary is considered as a data structure which contains a list of dictionary entries for a given set of terms, and a lookup mechanism which, given a source-language query term, consults this data structure to obtain a bag of one or more possible translations or equivalents of the term in question. An entry in a machine-readable bilingual dictionary is a data structure within a dictionary containing all of the necessary information for a given spelling of a source-language query term.

3. DATA MINING TECHNIQUES FOR CROSS LANGUAGE

Cross language information retrieval involves basically three problems [3]:

i. Crossing the language barrier i.e. Find the way to translate the term expressed in one language might be written in another.
ii. Determining which of the translations method to be choose. Selecting more than one translation methods helps in recall.
iii. Deciding the proper weight, if more than one translation is chosen.

Data mining automates the process of sifting through historical data in order to discover new information. Data mining techniques can yield the benefits of automation on existing software and hardware platforms to enhance the value of existing information resources, and can be implemented on new products and systems as they are brought on-line. The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning.

A data mining operation is achieved using one of a number of techniques or methods. Each technique can itself be implemented in different ways, using a variety of algorithms. Some of the classification algorithms that can be used on cross-language information retrieval platform are:

(i) Neural Networks
(ii) Decision Tree
(iii) K-nearest neighbor
(iv) Naïve Bayesian
(v) Cluster analysis.

3.1 Neural Networks

Neural networks consist of links, nodes and these nodes are consisting of output values and input values. Neural network technique can be explained in terms of layered subsystem, where data received in multiple phases can be fed to the next layer as a single phase.

When a query is issued by the user, the subsystem processes it and assigns a keyword to it. Intermediate subsystem indexes the given query and compares with document index. Based on the comparisons, the database system retrieves the relevant documents as a result [4].

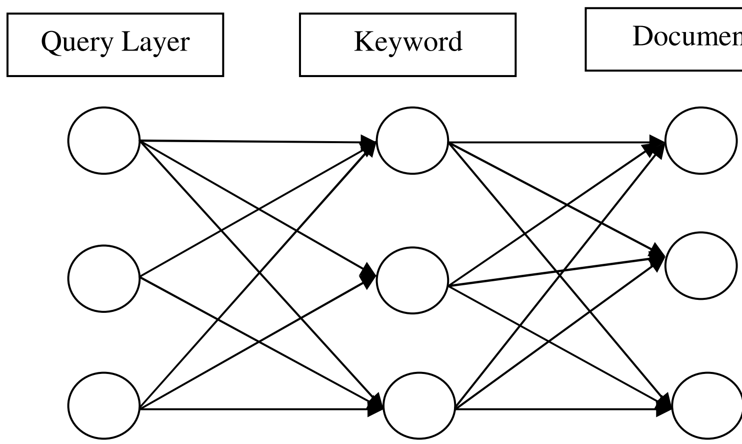

Figure3: Three layer neural network techniques for CLIR [4]

3.2 Decision Tree

A decision tree is a predictive model in which each branch can be viewed as classification of questions and leaves as partitions of the dataset with their classification. When a given object is subjected to a series of tests, in which the outcome contains class label to which object it has to be associated. In a decision tree, branch (non terminal) nodes are tests and leaf (terminal) nodes are class labels. Each branch node has number of child nodes [7].

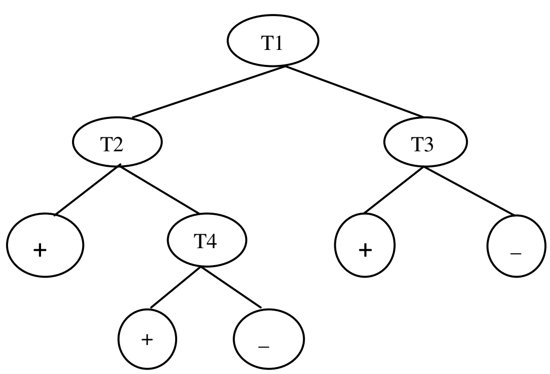

Figure4: Decision Tree for CLIR [7]

Classifications of an object in the decision tree begin from the root node and associated testes are applied. Depending on the obtained result an arc is traversed to the appropriate child node. If the subsequent node is a branch node, then its associated test is applied and arc is again traversed. This iterative process is continued until a leaf node is reached.

3.3 K-nearest neighbor

In K-nearest neighbor approach given a test document d, the system finds the K-nearest neighbors among training documents, and weight is assigned to the candidates using their classes. The total weight of the class is taken as the similarity score of each nearest neighbor document to the test document. If number of documents sharing the class is more in number, then per-neighbor weight of that class is added and the resulted weight is used as the likelihood score of that class with reference to the test document. A rank list can be obtained by sorting the scores of candidate classes. The decision rule in KNN classification is given as [10]:

$$\mathrm{score}(d, c_i) = \sum_{d_j \in \mathrm{KNN}(d)} \mathrm{Sim}(d, d_j)\delta(d_j, c_i)$$

Where

i. KNN (d) indicates the set of K-nearest neighbors of document d.

ii. (d_j,C_i) is the classification for document d_j with respect to class C_i, that is

$$\delta(d_j, c_i) = \begin{cases} 1 & d_j \in c_i \\ 0 & d_j \notin c_i \end{cases}$$

For test document d, it should be assigned the class that has the highest resulting weighted sum.

3.4 Naïve Bayesian

Naïve-Bayes technique is the organization of both predictive and descriptive. The conditional probability for each relationship is derived by analyzing the relationship between the dependent and independent variables. To generate a classification model only one pass through the training set makes this concept more efficient data mining technique. One of the drawbacks is that, it will not handle the continuous data, so independent or dependent variables that contain continuous values must be cased.

Prior probability is calculated by counting the number of occurrences of the dependent variable in the training dataset. Apart from this, naive-bayes is used to compute how frequently each independent variable value occurs in combination with each dependent variable value. These frequencies are then used to compute conditional probabilities that are combined with the prior probability to make the predictions.

3.5 Cluster Analysis

Clustering or Cluster analysis classifies the given data into groups (clusters) that are meaningful and useful. When a query is placed to the search engine, retrieved result may contain numerous pages. Clustering can be used to group these search results into a small number of clusters, each of which captures a particular aspect of the query. For instance, a query of “newspaper” might return web pages grouped into categories such as national new, international news, sports news, advertisements and entertainments. Each category (cluster) can be broken into subcategories (sub-clusters), producing a hierarchical structure that further assists a user’s exploration of the query results [13].

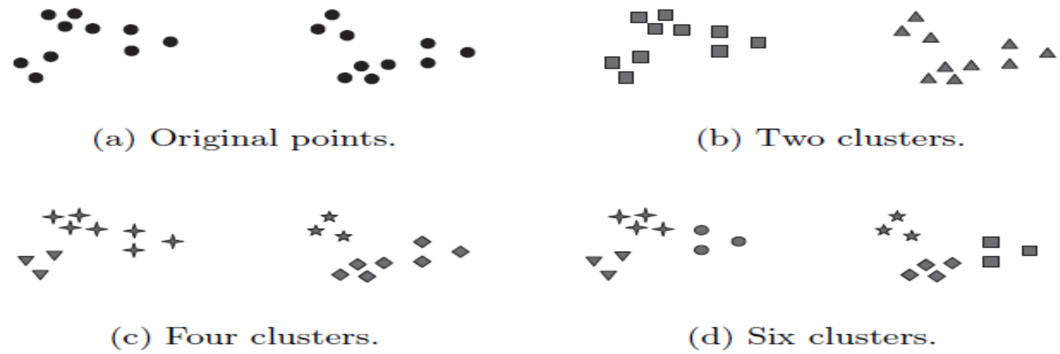

Figure 5: Clustering analysis for CLIR [13]

The following table [1] summarizes different types of algorithms that can be implemented for CLIR

	Techniques	Algorithms
1	Neural Networks	i. Perceptron ii. Backpropgation
2	Decision Tree	i. Quick reduct ii. Rough set based decision tree ensemble (RSDTE)
3	K-nearest neighbor	i. Neighbor-weighted K-nearest neighbor
4	Naïve Bayesian	i. Naïve Bayes Metiore (NBM) ii. Weighted Naïve Bayes Metiore (WNBM)
5	Cluster analysis	i. Hierarchical Agglomerative ii. Clustering without a recomputed matrix

Table 1: list of algorithms for different techniques

4. CONCLUSION

It has been observed Cross-Language retrieval is an important tool for several decades. But in the later period i.e. during1990s there is major focus and attention given in the research community so as to find out the user side issues of the tool. In today's scenario we have a quite few solutions available for handling such issues that are applicable to the Cross-language information retrieval, to mention translation ambiguity, lack of translation resources, and untranslatable terms. Many different ways of bridging the gap between query and document language have been devised and tested.

Predominately the technical related issues pertaining to the tool have been given more importance and researched extensively, and subsequently usability of that research has been mostly overlooked. There has been substantial information by looking at the statistics of languages available on the internet that there might be very huge market for cross-language products. From user perspective what he requires from such a system has been the focus of investigation.

Perhaps this paper has looked only at CLIR in terms of data mining techniques; there are variations of this being researched as well. Among them are cross-language multimedia, cross-language question answering, cross-language filtering, cross-language topic detection, cross-language summarization.

5. REFERENCES

[1]. Oard and Dorr, "A Survey of Multilingual Text Retrieval", *Technical Report UMIACS-TR-9619*, University of Maryland, 1996

[2]. Douglas W. Oard, "Alternative Approaches for Cross-Language Text Retrieval," in Cross-Language Text and Speech Retrieval, AAAI Technical Report SS-97-05.

[3]. Grefenstette, G., *Cross-Language Information Retrieval*. 1998, Boston: Kluwer Academic Publishers. 2000.

[4]. Igor Mokris, Lenka Skovajsova " Neural network model of system for information retrieval from text documents in Slovak language". Acta Electrotechnica et informatica no-3 vol 5,2005

[5]. Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20th ACM SIGIR conference on research and development in information retrieval (pp. 84–91).

[6]. Chu, H. (2003). Information representation and retrieval in the digital age. Medford, NJ: Information Today.

[7]. Ronnie Fanguy, Vijay Raghavan " Generating rule-based tree from decision tree for concept-based information retrieval", wss03 applications, products and services of web-based supported systems.

[8]. Soergel, D. (1985). Organizing information: Principles of database and retrieval systems**.** Orlando, FL: Academic Press.

[9]. Blair, D. C. (1990). Language and representation in information retrieval. Amsterdam: Elsevier Science.

[10]. Songbo Tan (2005). Neighbor- Weighted K-nearest neighbor for unbalanced text corpus: Elsevier Science.

[11]. Sigrid Roehling. Cross-Language information Retrieval.

[12]. Gina-Anne Levow, Douglas W Oard, Philip Resnik " Dictionary-based techniques for cross-language information retrieval" Elsevier, information processing and management 41 (2005) 523-547

[13]. Cluster Analysis: Basic Concepts and Algorithms. Page no-491.

[14]. Information Retrieval the scope of IR. HCIEncyclopediaIRShortEForDS. Page no-1.

Authors

Ms.Archana.M, is presently working as Assistant Professor in Department of Mater of Computer Application, R.V.College of Engineering, Bangalore, India, She has completed her MCA from Gulbarga University, MPhil from Alagappa University and has 5 years of teaching experiences. She has published papers in national and international Conferences. She is also a member of IAENG.

Dr.SumithraDevi K.A, is serving as a Director of Master of Computer Applications Department at R.V.Collage of Engineering, Bangalore, India, she earned PhD in Computer Science and Engineering from the Avinashilingam University for Women, Coimbatore, INDIA and she draws a strong back ground in VLSI Partitioning CAD Tool. She has been awarded by many national and international awards. She is the lifetime member of IEEE, WIE, ISTE and CSI. She is a registered PhD guide under VTU and is guiding two research students and one research candidate under the funded project.